Trends in COVID-19 Infected Cases and Deaths Based on Parametric and Nonparametric Regression Models

RAJARATHINAM ARUNACHALAM (arrathinam@yahoo.com)
Manonmaniam Sundaranar University https://orcid.org/0000-0002-3245-3181
TAMILSELVAN PAKKIRISAMY
Manonmaniam Sundaranar University https://orcid.org/0000-0003-4097-8874
Ramji Madhaiyan
Manonmaniam Sundarnar University https://orcid.org/0000-0002-4491-3093

Research Article

Keywords: Adjusted R2, Jarque-Bera test, Run test, Kernel Smoothing, Nonparametric regression, Epanechnikov-kernel

Posted Date: August 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-821046/v1

TRENDS IN COVID-19 INFECTED CASES AND DEATHS BASED ON PARAMETRIC AND NONPARAMETRIC REGRESSION MODELS

[1]Tamilselvan, P, Rajarathinam, A[*2] and Ramji, M[3].
[1,2,3]Department of statistics, Manonmaniam Sundaranar University, Tirunelveli-627 012, India
E-mail: ptamil74@gmail.com , arrathinamsu@gmail.com and 7jipgm@gmail.com
[*]Corresponding Author: arrathinammsu@gmail.com, Tel.: +91-8220226545.

Abstract: The present investigation was carried out to study the trends in COVID-19 infected cases and deaths based on the parametric, exponential smoothing and non-parametric regression models by using COVID-19 cumulative infected cases and deaths due to infections The statistically most suited parametric models are selected based on the highest adjusted R^2, significant regression co-efficient and co-efficient of determination (R^2). Appropriate model is selected based on the model performance measures such as, Root Mean Square Error, Mean Absolute Error, Mean Absolute Percentage Error, assumptions of normality and independence of residuals. Nonparametric estimates of underlying growth functions are computed at each and every time points.

Keywords: Adjusted R^2, Jarque-Bera test, Run test, Kernel Smoothing, Nonparametric regression, Epanechnikov-kernel.

Conflict of Interest : Authors have no Conflict of Interest

I. INTRODUCTION

1.1 Preamble

The COVID-19 pandemic in 2019 has received much attention, as it affected most economies worldwide and resulted in uncountable deaths. Because no antiviral drugs or vaccines exist, the number of new coronavirus-affected cases has tremendously increased, and many people have died. The development of various methodologies to analyse these pandemic data has become a very important research area regarding the prediction of future corona virus cases.

1.2 Review of Literature

Jiang et al. (2000), by applying time series-based kinetic model for infectious diseases, obtained trends and short-term predictions for the transmission of COVID-19.

Al-Rousan and Al-Najjar (2020),studied the effect of various factors, such as, sex, region, infection mode and birth year, on recovered and deceased cases due to COVID-19, in the South Korean region.

Gondauri et al., (2020) for studying and analysing the correlation between the total numbers of COVID-19 cases and recoveries in different countries, the chain-binomial type of Bailey's model is employed . Most of the studies investigated COVID-19 cases based on various regression and time series models because these models are frequently applied to examine the growth or trend of diseases.

Katris (2021) aimed to generate a time series-based procedure to track outbreaks. In the first stage, he used univariate time series models to present the evolution of the reported cases. He also used combinations of the models to provide more accurate and robust results and considered statistical probability distributions to generate future scenarios. The final step was

to build and use an epidemiological model (the time series susceptible-infected-recovered [tsiR] model) and to calculate the epidemiological ratio (R_0) to estimate the termination of the outbreak. In addition to feed-forward artificial neural networks and multivariate adaptive regression splines from the machine learning toolbox, the time series models deployed included the classical exponential smoothing and ARIMA approaches. The combinations included the simple mean, Granger-Newbold and Bates-Granger approaches.

Kumar and Roy, (2020) by using the Bailey's model with secondary data, calculated the removal rate, which is the percentage of removed persons in the infected population. Further, regression analysis is performed to show the linear relationship of this indicator with total infection rates. Finally, they described how the model could be linked with decision making.

Mittal, (2020) carried out an exploratory data analysis with the aim of elaborating a statistical model to better understand COVID-19 in India by thoroughly studying the cases reported in the country through 22 April 2020. The results of the study showed the impact of COVID-19 in India at the daily and weekly level and drew parallels between India and neighboring countries as well as severely affected countries.

Ogundoun et.al., (2020), by adopting the ordinary least squares (OLS) estimator to measure the impact of travel history and contact with travelers on the spread of COVID-19 in Nigeria and created forecasts by extracting data spanning 31, 2020, to May 29, 2020, from the Nigeria Centre for Disease Control (NCDC) website. The model assessed the period before and after travel restrictions were enforced by the federal government of Nigeria. The fitted model fit the dataset well and was free of any validity violations based on the diagnostic checks conducted. The results show that the government made the right decision in enforcing travel restrictions, with travel history and contact with travelers found to increase the chances of people being infected with COVID-19 by 85% and 88%, respectively; the authors concluded that the government should enforce this policy to contain COVID-19.

Nesteruk, (2020), used simple mathematical model predicted the characteristics of the epidemic caused by corona virus in mainland China. The optimal values of the SIR model parameters are identified with the use of statistical approach. The numbers of infected, susceptible, and removed persons versus time are predicted and compared with the new data obtained after February 10, 2020, when the calculations are completed.

Rajarathinam and Tamilselvan, (2021) studied the short- and long-term cointegration relationships between the cumulative number of COVID-19 infections and the cumulative numbers of deaths due to COVID-19 are studied by employing an autoregressive distributed lag model and bound cointegration tests. The stability of the estimated model is also assessed. The cumulative sum of the recursive residuals test and the cumulative sum of recursive residuals squares tests are used to assess the consistency of the model's parameters.

Takele (2020) used stochastic modelling to predict COVID-19 prevalence patterns in East African countries, mainly Ethiopia, Djibouti, Sudan, and Somalia. The study results showed that in the four months following June 30, 2020, the number of COVID-19-positive people in Ethiopia could rise from 5,846 to 56,610 in the average rate scenario.

1.3 Objectives of the study

Based on the above information, the present study aimed to assess the trends in the number of cases related to COVID-19, i.e., whether the number of cases increases or decreases, based on

the the parametric, exponential smoothing and nonparametric regression models by using COVID-19 cumulative infected cases and deaths due to infections data set.

II. MATERIALS AND METHODS

2.1. Materials

The cumulative total numbers of COVID-19 infections and deaths as of 31st June 2021, starting on 9th March 2020, were collected from the official website maintained by the Health and Family welfare department, Government of Tamil Nadu.Stat graphics Centurion XVI Ver.16.1.12 was used to estimate the model parameters, error diagnostics and to study the stability of the estimated model.

2.2. Methods

In parametric models different linear models viz., Linear trend, Quadratic trend, Exponential trend, S-Curve trend (Montgomery, et.al., 2003); different smoothing viz., Simple Exponential Smoothing, Brown' linear exponential Smoothing, Holt's Linear Exponential Smoothing (Makridakis, et al., 1998) ; different ARIMA time-series models (Box and Jenkins, 1976) and Nonparametric regression have been employed. The statistically most suited parametric models are selected on the basis of highest adjusted R^2, significant regression co-efficient, assumptions of residuals (normality and randomness). The performances of the different models have been carried out based on model performance measures such as (Root Mean Square Error) RMSE, (Mean Absolute Error) MAE and Mean Absolute Percentage Error (MAPE) values. Nonparametric estimates of underlying growth functions are computed at each and every time points. Residual analysis is carried out to test the randomness as well as normality. A relative growth rate is calculated based on best fitted models.

2.2.1. Nonparametic Regression (Hardle, 1990; Takezawa, 2006)

Nonparametric regression technique for functional estimation has become increasingly popular as a tool for data analysis. The technique imposes only few assumptions about shape of function and therefore it may be more flexible than usual parametric approaches. In many situations, we may not know the exact functional form and sometimes there may not be any parametric functional form to represent the data. In such situations, the nonparametric technique, which entirely depends on the data will be more suitable. This method is based on the local regression smoother and only assumption about the form of trend. The nonparametric techniques having comparable merit which broadly indicates the direction of the growth rates, and the usual statistical analysis, which provides both direction and dimension of growth rates.

2.2.2. Estimation of trend and growth rate (Jose et.al., 2008)

The nonparametric regression model with the additive error of the form $Y_i = m(x_i) + \varepsilon_i$, $x_i = i/n$, i=1,2,3, … , n where Y_i is the observation of the i^{th} time point, m is the trend function, which is assumed to be smooth, and ε_i are random errors with mean zero and finite variance $\sigma^2 < \infty$. The kernel weighted linear regression smoother (Fan, 1992) is used to estimate the trend function nonparametrically. The value of the local linear regression smoother at time x is the solution of a_0 to the following weighted least squares problem:

$$\sum_{i=1}^{n}\left[y_i - a_0 - a_1((x-x_i)/h)\right]^2 K_h((x-x_i)/h)$$

where K is a bounded symmetric kernel density function and h is the bandwidth. Let $\hat{a}_0$ and $\hat{a}_1$ be the solutions to the weighted least squares problem. The estimate of the trend function m(t) is given by $\hat{m}(t) = \hat{a}_0 = \sum_{j=1}^{n} W_{tj} y_j$

where

$$W_{tj} = \frac{K_j[s_2 - (x-x_j)s_1]}{s_0 s_2 - s_1^2} \quad K_j = K\left[\frac{x-x_j}{h}\right] \text{ and } s_l = \sum_{K=1}^{n} K\left(\frac{x-x_k}{h}\right)(x-x_k)^l$$

The optimum bandwidth h can be obtained by the method of cross-validation. The slope $m^{|}(x)$ of m(x) can be considered as the simple linear growth rate at the time point x. The estimate of $m^{|}(x)$ is given by $\hat{m}^{|}(x) = \hat{a}_1 = \sum_{j=1}^{n} W_{tj}^{|} y_j$ where $W_{tj}^{|} = \frac{K_j\left[(x-x_j)s_0 - s_1\right]}{s_0 s_2 - s_1^2}$

Under the assumption that the trend function m is smooth and $m(x) \neq 0$ for all $x \in [0, 1]$, the value of the relative growth rate at time X can be written as: $r_x = \frac{m^{|}(x)}{m(x)}$ .Since $\hat{m}(t) \to m(t)$ and $\hat{m}^{|}(x) \to m^{|}(x)$, a consistent estimate of the relative growth rate r_x is given by:

$$\hat{r}_x = \frac{\hat{m}^{|}(x)}{\hat{m}(x)} \to r_x$$

Taking arithmetic mean, the requisite compound growth rate over a given time-period may be obtained.

3.1. Descriptive Statistics

Descriptive statistics have been calculated to know the nature of the distribution of the study variable and reported in the Table 1. The results reveal that both the variables the Jarque-Bera statistics p-values are found to be significant indicating that the study variables are not normally distributed. The maximum number of COVID-19 infected cases are 532529 registered at Chennai and minimum is 11056.00 registered at Perambalur; the maximum cumulative number of COVID-19 deaths due to infections is 8187.000 registered at Chennai and minimum is 164.0000 registered at Nilgiris.

Table 1: Descriptive Statistics of Variables

	CASES	DEATH
Mean	65189.16	858.3421
Median	43500.00	510.5000
Maximum	532529.0	8187.000
Minimum	11056.00	164.0000
Std. Dev.	87791.15	1326.613
Skewness	4.266072	4.644745
Kurtosis	22.61554	25.94028
Jarque-Bera	724.4809	969.8724
Probability	0.000000	0.000000

Sum	2477188.	32617.00
Sum Sq. Dev.	2.85E+11	65116403
Observations	38	38

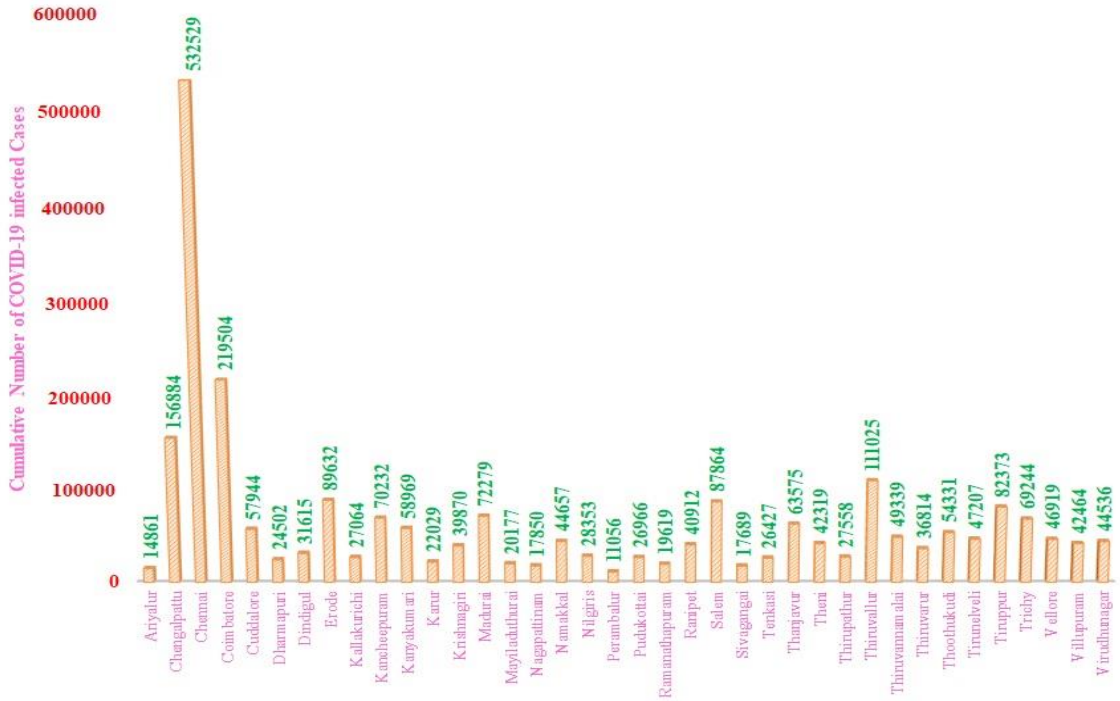

Fig.1. District wise cumulative number of COVID-19 infected cases

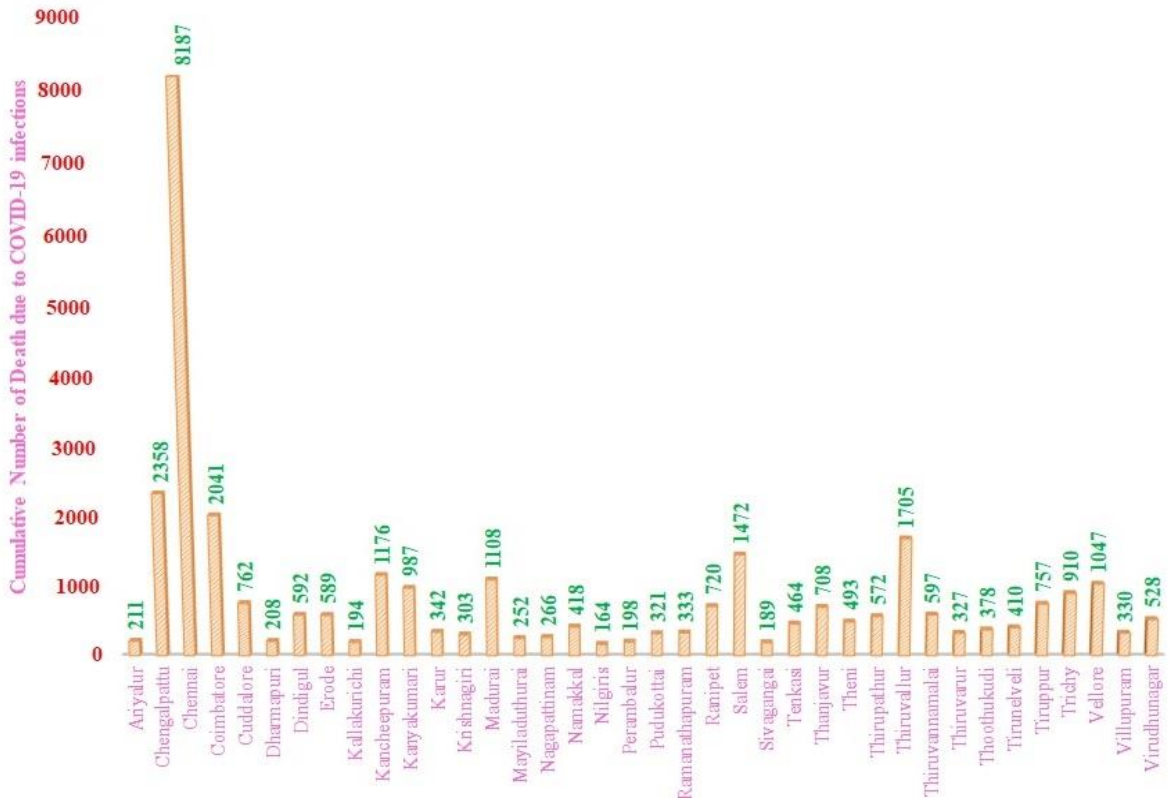

Fig.2. District wise cumulative number of deaths due to COVID-19 infection

Table 2: Characteristics of fitted parametric models for COVID-19 infected case

Model	Model Performance Measures		
	RMSE	MAE	MAPE
Linear trend	84902	47688	120.89
Quadratic trend	79662	39475	95.10
Exponential trend	90103	38515	65.36
S-curve trend	90653	38031	63.93
Simple exponential	86840	47311	122.59
Brown's linear exponential	91004	49821	123.54
Holt's linear exponential	88807	50289	117.81
ARIMA(1,1,2)	69639	40661	99.63
ARIMA(2,1,0)	73371	41081	100.46
ARIMA(2,1,1)	74360	43496	100.81
ARIMA(2,1,2)	72708	39135	101.40
ARIMA(0,1,2)	77098	39701	78.72

3.3. Trends in cumulative number of deaths due to COVID-19 based on parametric model

The results presented in Table 3 reveals that among the parametric model, the ARIMA (2,1,0) time series model has the lowest values of RMSE (1221.02), MAE (608.53) and MAPE (148.83). Hence among the parametric models the ARIMA (2,1,0) is found suitable to fit the trends in in the cumulative number of death due to COVID-19 infections.

Table 3: Characteristics of fitted parametric models for the COVID-19 cumulative deaths

Model	Model Performance measures		
	RMSE	MAE	MAPE
Linear trend	1287.97	662.48	138.33
Quadratic trend	1230.53	590.77	119.35
Exponential trend	1363.75	553.68	71.70
S-curve trend	1368.64	553.97	71.61
Simple exponential	1332.77	748.25	160.44
Brown's linear exponential	1379.91	815.73	180.58
Holt's linear exponential	1363.45	755.80	143.96
ARIMA(2,1,0)	1221.02	608.53	148.83
ARIMA(0,1,2)	1280.75	605.08	100.68
ARIMA(2,1,1)	1249.81	696.41	142.31
ARIMA(0,1,1)	1326.34	729.04	147.20
ARIMA(1,1,1)	1301.91	580.79	88.79

3.4. Trends in cumulative numbers COVID-19 infected cases based on nonparametric regression model

The nonparametric regression model is employed to fit the trends in number of COVID-19 infected cases. Nonparametric estimates of underlying growth function are computed at each and every time points. Residual analysis showed that the assumptions of independence of errors are not violated at 5% level of significance. The RMSE, MAE, MAPE values are 62090.77, 31047.79 and 65.78, respectively. These values are found to be much lower than that of obtained through the parametric models, indicating thereby the superiority of this approach over the parametric approach. Nonparametric regression model is selected as the best fitted trend function for the number of COVID-19 infected cases and depicted in the fig.3.

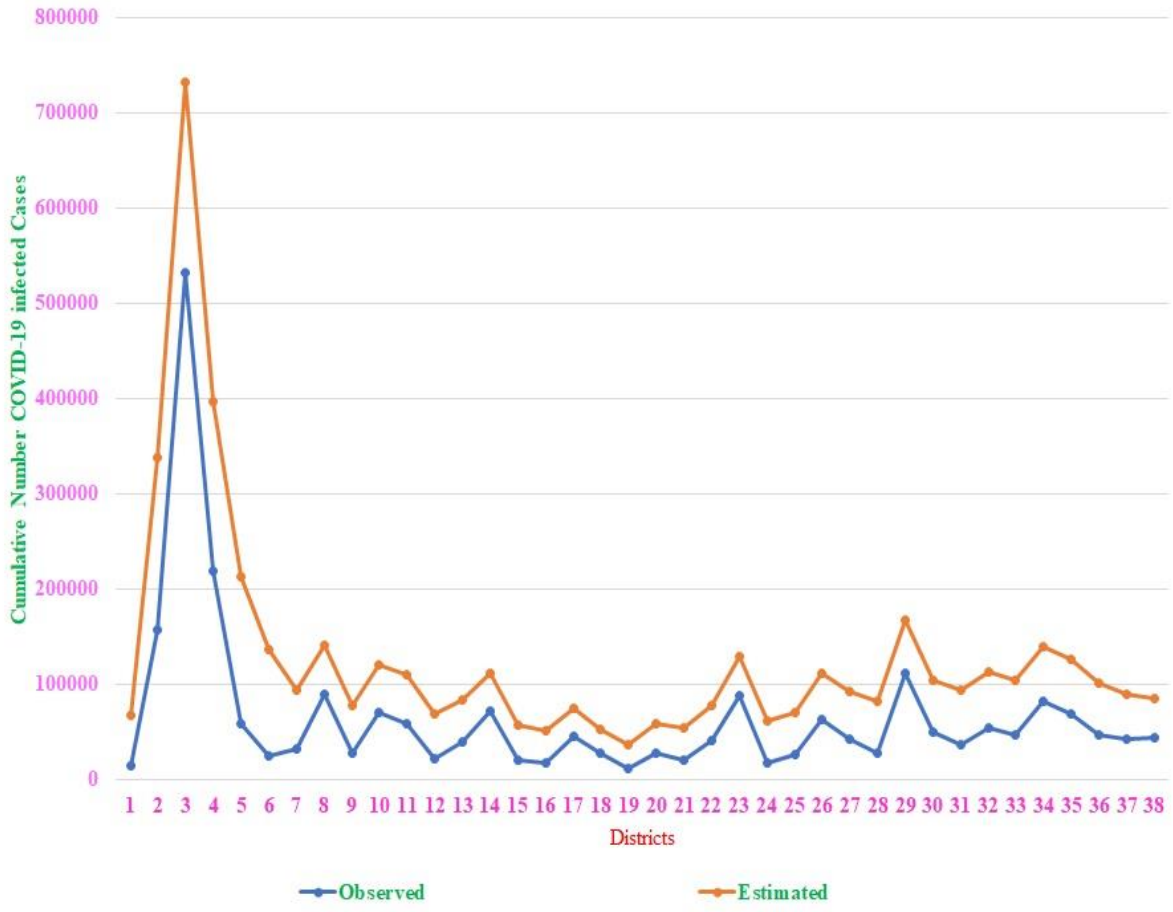

Fig.3: Trends in cumulative number of COVID-19 infected cases

3.5. Trends in cumulative numbers of deaths due to COVID-19 based on nonparametric regression model

The nonparametric regression model is employed to fit the trends in cumulative numbers of deaths due to COVID-19. Nonparametric estimates of underlying growth function are computed at each and every time points. Residual analysis showed that the assumptions of independence of errors are not violated at 5% level of significance. The RMSE, MAE, MAPE values are 980.47, 464.70 and 78.49, respectively. These values are found to be much lower than that of obtained through the parametric models, indicating thereby the superiority of this approach over the parametric approach. Nonparametric regression model is selected as the best

fitted trend function for the number of COVID-19 infected cases and depicted in the following fig.4.

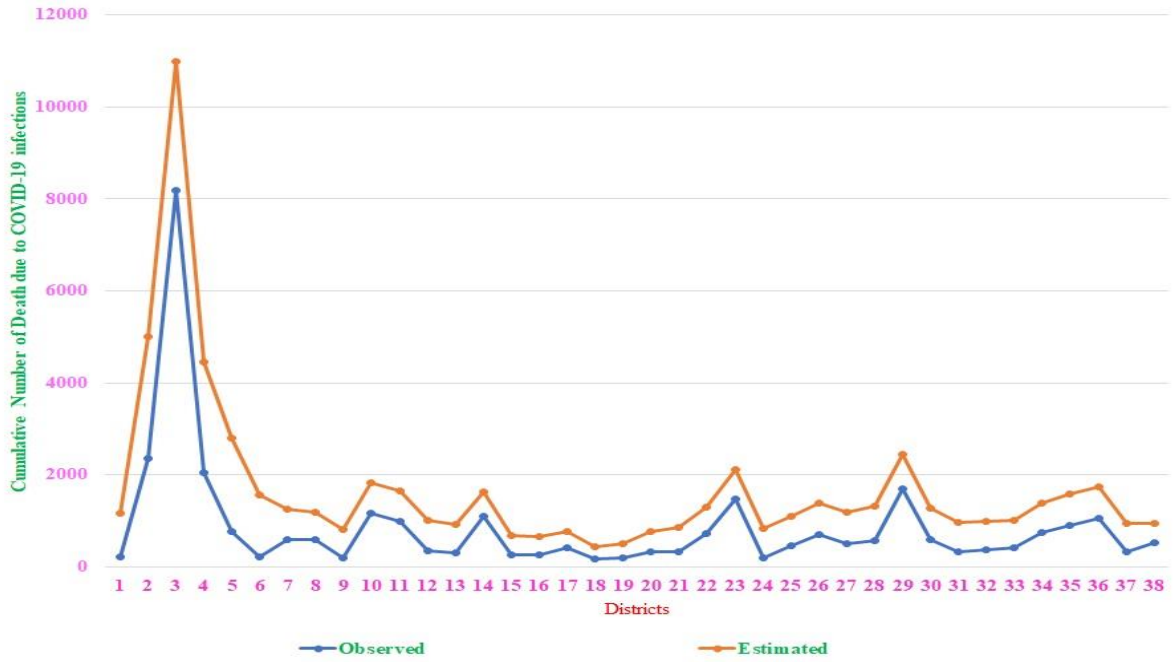

Fig. 4: Trends in cumulative number of deaths due to COVID-19 infection

III. CONCLUSION

Results reveal that none of the parametric models have been found suitable to study the trends in cumulative number of COVID-19 infected cases and number of death due to COVID-19 infections.

ACKNOWLEDGMENT

Authors are thankfully acknowledged the editor and anonymous reviewer for their valuable suggestions and comments to improve the earlier version of this paper.

References

1. Al-Rousan, N., and H. Al-Najjar. Data Analysis of Coronavirus COVID-19 Epidemic in South Korea based on Recovered and Death Cases. *Journal of Medical Virology*, 92 (2020), pp. 1603-1608.
2. Baltagi, B.H. *Econometric Analysis of Panel Data*, (2001), Wiley, New York, NY.
3. Box,G.E., and Jenkins, G.M. *Time Series Analysis : Forecasting and Control*, (1976), San Francisco : Holden-Day.
4. Breusch, T.S., and A.R. Pagan. A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, 47 (1979), pp. 1287-1294.
5. Chatterjee, A., M.W. Gerdes, and S.G. Martinez. Statistical Explorations and Univariate Time-series Analysis on COVID-19 data sets to understand the trend of Disease Spreading and Death. *Sensors (Basel)*, 20 (2020), 3089.
6. Godfrey, L.G. Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables. *Econometrica*, 46 (1978), pp.1293-1301.

7. Gondauri, D., E. Mikautadze, and M. Batiashvili. Research on COVID-19 Virus Spreading Statistics based on the Examples of the Cases from Different Countries. *Electron Journal of General Medicine*, 17 (2020), pp.1-4.
8. Gujarati, D.N., D.C.Porter, and G.Sangeetha. *Basic Econometrics*, 5th edition. McGraw Hill Education (2017), New York, NY.
9. Hardle, W. *Applied Nonparametric Regression*. (1990), 1st Edn., Cambridge University Press, New York, USA.
10. Hadri, K. Testing for Units Roots in Heterogeneous Panel Data. *Econometrics Journal*, 3 (2000), pp. 148-161.
11. Hausman, J.A. Specification Tests in Econometrics. *Econometrica*, 46 (1978), pp. 1251-1271.
12. Hsiao, C. *Analysis of Panel Data*. Cambridge University Press (2003), Cambridge.
13. Jiang, X., B. Zhao, and J. Cao. Statistical Analysis on COVID-19. *Bio-medical Journal of Scientific & Technical Research*, 26 (2020), pp. 19716-19727.
14. Levin, A., C.F. Lin, and C.S.J. Chu. Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties. *Journal of Econometrics*, 108 (2002), pp. 1-24.
15. Fan, J. Design Adaptive Nonparametric regression. *Journal of American Statistical Association*, (1992), 87, pp.998-1004.
16. Makridakis, S., Wheelwright, S.C., and Hyndman, R.J. *Forecasting: Methods and Applications*, J. Wiley and Sons (1998), New York.
17. Mittal, S., An exploratory data analysis of COVID-19 in India. *International Journal of Engineering and Technical Research*,2020, 9, 580-584.
18. Montgomery, D.C., Peck, E.A., and Vining, G.G. *Introduction to Linear Regression Analysis*, John Wiley & Sons (2003), Inc.
19. Rajarathinam, A., and P.Tamilselvan, Autoregressive Distributed Lag Model of COVID-19 Cases and Deaths, *Appl. Math. Inf.* Sci,2021 (in press).
20. Nesteruk, I. Statistical-Based predictions of coronavirus Epidemic Spreading in Mainland China. *Innov.Biosyst. Bioeng*, 4 (1) (2020), pp.13-18.
21. Takele, R. Stochastic modelling for predicting COVID-19 prevalence in East Africa Countries. *Infectious Disease Modelling*, 5 (2020), pp. 598-607.
22. Takezawa, K. *Introduction to Nonparametric Regression*, (2006), John Wiley & Sons.
23. Wald, A. Tests of Statistical hypothesis concerning several parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society*, 54 (1943), pp. 426-482.